DOCUMENT RESUME

ED 185 563 CS 205 478

AUTHOR Lutz, William D.
TITLE How to Read 55,000 Essays a Year, and Love It.
PUB DATE Mar 80
NOTE 9p.: Paper presented at the Annual Meeting of the Conference on College Composition and Communication (31st, Washington, DC, March 13-15, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Basic Skills: *Evaluation Methods; Higher Education; Program Descriptions; State Programs; Student Evaluation; Student Placement; *Test Construction; *Testing Programs; *Writing (Composition)
IDENTIFIERS Holistic Scoring; New Jersey; *Writing Evaluation

ABSTRACT

The development of the portion of a basic skills test that would adequately assess the writing ability of entering college freshmen in New Jersey is described in this paper. The stages of the test development process that are discussed include the decision to incorporate a writing sample in the form of an essay on a specified topic, the selection and piloting of an essay topic, the recruitment and training of readers to holistically score the essays, and a statistical analysis of the writing test data. Listed as significant results of this effort are the provision of a data base for a statewide assessment of the writing ability of all freshmen entering all colleges in the state, and the bringing together of college faculty from throughout the state for scoring essays and interpreting test results. (AEA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
William D. Lutz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED185563

HOW TO READ 55,000 ESSAYS A YEAR, AND LOVE IT

William D. Lutz
English Department
Rutgers University
Camden, NJ 08102

In 1977 the State of New Jersey decided to assess the reading, writing, and mathematics skills of all freshmen entering all New Jersey public colleges and universities. To administer this testing program, the New Jersey Board of Higher Education appointed the Basic Skills Council, a group of ten faculty from colleges across the state. The Council in turn was empowered to appoint an executive director to serve as its administrative officer. I served as the first Executive Director of the New Jersey Basic Skills Council from 1977 to 1979 while I was on leave of absence from my position as Director of Freshman English at Rutgers University in Camden.

The first task confronting the council was to develop a test which would adequately assess the reading, writing, and mathematics skills of entering freshmen. The results of the test were to be used to identify those students who would need remedial/developmental assistance before undertaking regular college work. Colleges were required to place such students into appropriate remedial/developmental courses once their need for remediation determined.

After much research and discussion, the Council decided to develop its own test and not to use any already published tests. The Council also made another important decision. If writing were to be assessed, it would be assessed only if a writing sample were included in the assessment process. The implications of this decision were, I think, apparent. Each freshman student entering a New Jersey public college would have to write an essay and that essay would have to be scored by faculty readers. At the time this decision was made, we really weren't sure how many students we were talking about. The best guesstimate available on the number of students who would be tested was 55,000. There were momentary visions of the ten members of the Council, and the Director,

CS205478

locked in a rather large room reading 55,000 essays.

But the problem was real. How would we go about reading such a large number of essays? One important restriction on any decision the Council made was that the essays would have to be scored according to a statistically reliable procedure. That is, we simply could not allow each college to score its own essays and submit the results to the Council. The variations in the scores coming from the faculty at 38 different colleges simply would not contribute to statistical reliability, nor would such variations contribute very much to score consistency. The logistics involved in reading so many essays was a major problem. But an even more important problem was cost. How much would it cost to read all those essays? More importantly, how would we ever convince those who control the purse strings to give us enough money to read those essays, especially since most people believe that multiple-choice tests are just fine for testing anything, including writing?

The Council began by reaffirming its belief that any assessment of writing must include a writing sample that is reliably scored. Taking a cue from Godshalk's THE MEASUREMENT OF WRITING ABILITY, the Council decided to use a multiple-choice test and a writing sample. The Council would develop the multiple-choice test so that the results of that portion of the test could be combined with the scores on the essay to produce one score that would give a broad picture of a student's writing ability. This approach is consistent with the findings of Godshalk. And so the work began.

A multiple-choice test was developed in cooperation with Educational Testing Service. A special faculty committee of reading and writing instructors from New Jersey colleges worked with test development specialists from ETS to develop a multiple-choice test with which the instructors were comfortable. The essay portion of the test would consist of a twenty minute essay written on an assigned topic. The essays would then be read holistically by faculty

from New Jersey colleges who would gather for the prupose of spending a wonderful weekend reading a few thousand essays. The first holistic essay reading for the New Jersey testing program began about six months before a reader even picked up an essay to score.

A special faculty committee of reading and writing instructors gathered about six months before the first scheduled reading. The committee's purpose at this meeting was to review a list of some 125 essay topics which had been submitted by experienced college writing instructors. From this list of topics the committee was eventually to come up with the one essay topic that would be used in the testing program during the coming academic year. For a full day these faculty examined the suggested topics and evaluated each one. Could a student write on this topic immediately? Did the topic call for any special background knowledge? Would the topic call for a negative essay that might prejudice the reader? Was the topic interesting enough so that a reader wouldn't go stark raving mad after reading about a thousand essays? And so it went. Each topic was ruthlessly examined, challenged, defended. at the end of the day, 10 topics remained, and two of those were written during the meeting and weren't even on the original list. But at least there were ten.

Now the committee was ready to go to the next step -- test the topics. All ten topics were given to college freshmen across the state and the resulting essays read. Once again the committee gathered for a day, only this time they spent the day reading a few thousand essays and discussing whether the various topics worked. Could students write on the topic? Did the topic elicit good essays as well as bad? Did the topic allow a variety of approaches, or did it seem to require the same approach from each student? Was the topic flexible enough to allow students to demonstrate what they could do, or did the topic unduly restrict students? Did the topic elicit negative essays; for example, a series of essays against women's liberation? And so the essays were read

and the topics evaluated. At the end of the day one topic had survived.

From a beginning of a list of 125 topics we now had one that we felt fairly confident we could use. By the way, the cost of developing this topic was approximately $25,000. And we hadn't assessed the writing ability of one student yet, nor did this figure include the cost of developing the multiple-choice test. But at least we had an essay topic which had been developed by New Jersey college faculty and which had been pre-tested on New Jersey college freshmen. We were making progress.

The next step was to administer the test. Each college had the responsibility for administering the test to its freshmen after they had been admitted. All tests were submitted to the Basic Skills Council for scoring and score reporting. It was a simple matter to score the multiple-choice tests. After a while, it became a relatively simple matter to score the essays.

The essays were read according to an established schedule. That is, the colleges could administer the tests whenever they wanted and submit the tests to the Council by a certain date for scoring. From May, 1978 to February 1979 there were nine essay reading sessions which scored the essays written by all the freshmen admitted for the fall, 1978 semester and for those admitted for the spring, 1979 semester. During that time, 54,060 essays were read twice.

An essay reading really begins, as I have said, when the faculty committee gathers to begin work on writing topics for pre-testing. After the topics have been pre-tested and a single topic selected for use in the testing program, the next step is to administer the test. Then, when you have 12,000 or so essays sitting around you are ready for the next step in the holistic scoring procedure.

The chief reader who was responsible for the essay reading sat down with the assistant chief reader and spent a long day reading a few hundred essays from among those submitted for scoring. It was his (or her) responsibility to

select sample essays that properly represented the universe of essays to be scored during the reading. That is, the chief reader looked for essays which illustrated the range of essays to be scored, from the best to the worst, and any that illustrated special problems that might be encountered in this batch of essays.

The day before the reading, the table leaders met with the chief and assistant chief readers. During this meeting, which usually lasted a good half day, the sample essays were read and discussed. The group came to an agreement on which essays in their judgment best illustrated the various scores that would be used in the essay reading. We used a four-point scale. (Notice it is an even number with no mid-point.) The group not only agreed upon which sample essays represented a 4 (good) essay and which a 1 (poor), but also which essays illustrated a 3/2 split; that is, an essay which could be considered a low 3 by one reader and a high 2 by another, a legitimate difference of opinion and allowable in the scoring system. These and other possible scoring variations were discussed and illustrative essays selected.

Early the next morning (usually 8 o'clock on a Saturday), the faculty readers gathered. The readers had been selected from a pool of applicants, with one out of three or four applicants chosen. The faculty came from two and four year colleges across the state. Over sixty faculty may have been needed for any one essay reading. All had their mileage expenses paid to drive to and from the central meeting place, all had their room and board paid for the weekend, and all received an honorarium for their services. Conducting a holistic reading is not inexpensive.

The scoring session began with the chief reader distributing a set of sample essays which each reader scored. All the scores were reported and the chief reader reported how the table leaders scored the essays. The chief reader then determined if the group was reading on scale, or if there were disagreement over the scoring of any essay. Did all of the readers give sample

essays a 3, or were there some who gave it a 2 or perhaps even a 1? If so, then there would have to be some discussion on that essay. This process was repeated until everyone was scoring on scale; that is, the readers were in agreement over which essays should receive which score. Once this stage had been reached (and only the chief reader could decide when it had been reached), the readers began reading for real. Frequently during the reading the chief reader repeated the process of distributing packets of sample essays and going through the process of setting the scale to ensure that no one was tending to read too high or too low. Also during the reading the table leaders would randomly select an essay which a reader had already scored, score it himself, and then compare the two scores to see if there were any great variation. During the reading for the entire two days readers were constantly monitored to ensure that they continued to score consistently.

Each essay was read twice by two different readers. The essays had been randomized so that no reader received a packet of essays from the same college. After the essays had been scored by a reader, they were taken to another room where they were once again randomized and given to another reader. Thus, two readers did not simply exchange packets of essays. Also, the system was so designed that no reader ever read the same essay twice. This system of randomization ensured the widest possible distribution of essays among all the readers.

At the conclusion of the reading, the scores on the essays were read by a computer and merged with the scores on the multiple-choice test. The computer also produced a statistical analysis of the essay reading. It gave such information as how many essays an individual reader read (the reader's rate), whether the reader tended to read high, low, or on scale, whether the reader produced a lot of splits (an essay with two scores more than a point apart), and what percentage of the reader's scores were in complete agreement with the second reader's scores. There were also group data available on the reading.

The reliability of the reading, the percentage of agreement between readings, and the percentage of splits were some of the more interesting and useful statistics provided by the computer analysis.

For the very first essay reading we conducted, 6,188 essays were read with a reading reliability of .77, a percentage of agreement between readings of 55%, and a split percentage of 2.4%. All in all, not a bad performance. But as time went on, and we read a few thousand more essays and our readers became more experienced, our statistical profile for the readings also improved greatly. By the ninth essay reading the reading reliability was .82, the percentage of agreement between readings was 62%, and the percentage of splits was just 1%.

The results of all this effort have been more than worth the time and money expended. Although it costs approximately $2.50 to score each essay, the Basic Skills Council believes it is absolutely essential that the essay portion of the testing program be retained. First, we believe that the message should be sent to both students and teachers that any assessment of writing should include a writing sample. Second, the inclusion of the essay contributes to increasing the validity and reliability of the score produced by the writing test. Third, the essay has provided the data base for a statewide assessment of the writing ability of all freshmen entering all colleges in the state. Moreover, this data base is longitudinal since the test is administered to each entering freshman class. Fourth, the essay scoring sessions bring together faculty from colleges all over the state. For probably the first time these faculty are joined in a cooperative effort which allows them to meet and exchange ideas, as well as get some sense of the writing ability of students other than those at their colleges. Fifth, the faculty design and administer the testing program, including the selection of the essay topic. And sixth, the faculty interpret the results of the test for purposes of placement.

I am no longer the Director of the Basic Skills Council. I have served my two years and have returned to Rutgers. My replacement is another faculty

member who has taken a leave from his college to serve as Director. Since I have left the Council, an additional 63,586 essays have been scored for a two year total of 117,646. The results continue to be excellent. Beginning next academic year, the essays will be read on a six point scale. And work is about to begin on developing an essay topic for use during the 1981-82 academic year.

I look back now on my two years and have few if any regrets and only the best of memories for the time I administered a testing program during which 54,060 essays were read, and read reliability. It was a most pleasant experience; I could even say I loved it.

WDL/aml

Conference on College Composition and
Composition
Washington, D.C.
March 14, 1980